

# Advanced Computational Statistics – Spring 2025

## Assignment for Lecture 5

Frank Miller, [frank.miller@liu.se](mailto:frank.miller@liu.se),  
Department of Computer and Information Science, Linköpings University

April 29, 2025

Perform the solutions individually and send your report **until May 14** by email to me. Try to keep this deadline. However, if you have problems with it, there will be a final deadline on September 30 for all assignments. Please send me one **pdf-file** with your report and additionally, please send me your code in one separate **plain-text file**. **The following two problems are mandatory, but part c in the second is optional.**

### Problem 5.1

In the lectures, an EM algorithm was presented for the case of a univariate normal mixture model with two components; you can find also a text-file `emalg.txt` with the code in R, Python, Julia, and Matlab on the course homepage. Your task is here to generalize this EM algorithm to bivariate normal mixtures.

- Use the algorithm from the lecture as start and modify it for the case of **two-dimensional observations** which come from the mixture of two bivariate normal distributions; but allow for user specified starting values. Replace the stopping criterion with a stopping criterion based on the change in the expected log-likelihood.
- Use the dataset `bivardat.csv` on the course homepage which contains  $n = 1000$  observations. Create a two-dimensional point plot of the data. Make a choice for starting values of all model-parameters and discuss why you have chosen the starting values in this way.
- Fit the bivariate normal mixture model to the data using your program from a. and your choice of starting values from b. Check the convergence of all model-parameters.
- Check if results are depending on starting values by considering alternatives. If results differ: which are the better results and why?

## Problem 5.2

We consider again as in Problem 3.2 and 4.2 the experiment investigating how the growth of garden cress depends on a (potentially) toxic fertilizer. The data is on the course homepage below Topic 3 in the file `cressdata.txt` (columns: observation number  $1, \dots, 81$ , fertilizer concentration in %, yield in *mg*). Here, you are supposed to fit a quadratic regression,  $y = \beta_0 + \beta_1 x + \beta_2 x^2$ , where  $x$  = concentration and  $y$  = yield.

- a. Fit the quadratic regression and estimate the three coefficients  $\beta_0, \beta_1, \beta_2$  of the regression together with their 95%-confidence intervals using a standard function (e.g. the R-function `lm`). Create a plot for yield vs. concentration and add the estimated regression curve to the plot.
- b. Derive a 95%-bootstrap confidence interval for the three model parameters based on the percentile method (which was used in the lecture). Do not use a bootstrap package for this calculation (and not the function `bootstrp` in Matlab); program the bootstrap on your own. Use at least 10000 bootstrap replicates.
- c. **(optional)** Program your own  $BC_a$ -confidence interval and derive the 95%-confidence interval for  $\beta_2$ .
- d. Use now a bootstrap-package (e.g., `boot` in R, `scipy.stats.bootstrap` in Python, `Bootstrap.jl` in Julia, or the function `bootstrp` in Matlab) and derive the 95%-confidence interval for  $\beta_2$  using both the percentile- and the  $BC_a$ -method.
- e. Compare the different confidence intervals for  $\beta_2$ .